METRON
https://doi.org/10.1007/s40300-024-00276-5

Neither agree nor disagree: use and misuse of the neutral response category in Likert-type scales

Miloš Kankaraš[1] · Stefania Capecchi[2]

Received: 23 November 2023 / Accepted: 20 August 2024

Abstract

The Likert-type scales are among the most widely implemented instruments in social sciences, nonetheless, it is not clear so far whether such scales should or should not employ a mid-point "neutral" response option. While a mid-point category might improve the psychometric properties of survey instruments when appropriately applied, it has been argued that respondents often tend to use it in several invalid ways. This study aims to examine how a neutral response modality may influence the scales' psychometric properties. We conducted two types of survey experiments employing a between-subjects and a within-subjects design, comparing psychometric properties of twelve personality scales in both cases i.e., with and without the neutral response category. Our findings show that the scales presenting the neutral category allow to some extent for better psychometric characteristics, both in terms of their reliability and with respect to the proportion of accounted variance by the first factors. Results also suggest that most respondents seem to use the neutral category validly. However, there are also indications that a minority of respondents actually employ the neutral answer as an "escape" option, especially when asked socially sensitive questions.

Keywords Neutral response category · Likert-type scales · Survey experiment · Agreement scales · Personality · IRT

1 Introduction

Likert rating scales are among the most common forms of psychological assessment instruments [13, 23]. They are regularly utilized to allow individuals to express their standing on a given topic by choosing among several response categories ordered in between two extreme

✉ Stefania Capecchi
stefania.capecchi@unina.it

Miloš Kankaraš
m.kankaras@unesco.org

[1] UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development (MGIEP), New Delhi, India

[2] Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò, 80128 Naples, Italy

Published online: 26 September 2024

modalities. As opposed to binary questions admitting only two answer options, Likert-type questions allow for a more granular response and consequently for obtaining more reliable and valid measures of psychological concepts. Although the most prominent form of the Likert scale examines respondents' degree of agreement with given statements, these scales can also investigate the frequency, intensity, likelihood, importance, and interest as expressed by interviewees. Their versatility, ease of creation, application, and understanding, together with the quantifiable nature of captured information, make such scales suitable across a broad scope of academic and practical purposes [14, 16].

However, the use of Likert scales entails several unresolved issues and constraints. For example, it is still unclear which type of Likert scale is the most suitable for a given purpose. As a result, both agreement, importance, and frequency Likert scales are used to assess the same psychological constructs and personality traits [14]. Various forms of response styles and social desirability biases can substantially affect respondents' answers to Likert scales. Also, researchers constructing these scales are always confronted with how many answer options to offer, with rating scales varying between 3 to 10 to even 100 points. Another issue for consideration is whether a response scale should be unipolar in its wording structure (e.g., not satisfied vs. satisfied) or of a bipolar nature (e.g., agree vs. disagree). Besides, a pivotal dilemma arises concerning the presence/absence of a mid-point category, which is the topic of this study.

The typical rating scale applied in psychological assessment consists of between five and seven bipolar response options. In the case of an odd-numbered bipolar rating scale, the mid-point option usually indicates neutrality or ambivalence (e.g., "neither agree nor disagree"). On the other hand, the even-numbered Likert scales do not offer the mid-scale, "neutral" options to respondents, which is why they are sometimes also labelled as "forced-choice scales" when a response needs to be selected from one of the two opposing sides [1]. In the construction of the response scale, it is therefore essential to determine whether to use an even or an odd number of response categories.

From the theoretical measurement perspective, it is assumed that the mid-point option may genuinely represent "neutral" views. Such options are those selected if the respondent's construct standing is right between the positive and negative possibilities on the response continuum [19]. However, the main concern with the mid-point option in Likert scales is the chance that respondents do not employ it adequately, i.e., to indicate the moderate/uncertain standing on an item or non-applicable response [20].

Over the last decades, researchers have discussed and investigated several alternative ways in which the mid-point neutral response category may be interpreted and employed, each of which could jeopardize and attenuate the instruments' psychometric properties. Some respondents may select a middle alternative to minimize cognitive costs, and, even when they could, if pushed, they end up providing a directional response. Krosnick referred to this type of response as 'satisficing' [17, 18]. It could also be the case that some interviewees choose the mid-point option to give socially desirable answers—sometimes referred to as a "hidden don't know" response [32]. Likewise, to "please" the interviewer, respondents might pick the neutral category to avoid giving what they consider an unacceptable answer [12]. When replying to an unclear, intrusive, or introspective question, the neutral alternative could be seen as a tactic to avoid difficult choices [19]. Another plausible interpretation of the selection of the mid-point category on a Likert-type scale is the absence of a formulated opinion: respondents could use such a modality to hide their ignorance or lack of opinion, even when an actual "I don't know" option is provided. This response behaviour could also be linked to indifference or limited interest in the topic of the statement or reflect the dilemmatic nature of specific issues or disagreement with the question's terms [3].

In the study conducted by Nadler et al. [26], different response option formats were tested. Results show that participants choose the "neither" option for all the items in a 5-point response scale more frequently than the "no opinion" category provided along with a 4-point scale. Authors suggest that participants may have viewed the midpoint as more socially desirable than the "no opinion" answer when it comes to expressing their attitudes. The qualitative part of the study examines how the respondents interpret the given mid-point option: they applied various meanings to it, and their answers varied widely, but it was more commonly understood as "no opinion", "don't care", "unsure" or "neutral". When individuals are administered personality questionnaires and inventories, they tend to use the mid-point modality as a sort of "it depends" orientation, therefore suggesting that their response might be conditional and that certain items may require more extensive specification and/or contextualization [19].

A feasible solution to address this kind of problems is omitting the mid-point and including a "no opinion" response as an additional modality that should count as a missing value [20]. In some cases, when an even number of response options is provided, respondents may feel pressured to take either a positive or a negative stance on the scale. This circumstance could make it difficult to express neutral feelings: in such a way, a genuinely neutral attitude could not be conveyed and measured [3]. Respondents may be more likely to interpret the neutral category as a non-response when it is placed at the ends of the scale rather than in the middle. As a result, when the neutral option is offered separately from the other rating choices, the selection of the mid-point alternative tends to decrease [21].

When discussing the optimal number of response categories, some studies suggest that the reliability of Likert-type scales, as measured by the alpha coefficient, increases when more response options are provided. The same applies for factorial validity assessed through the percentage of variance explained by the first factor since it decreases as the number of response categories is reduced. Considering both criteria of reliability and validity, some scholars argue that a mid-point category is effective for improving these psychometric properties [7, 25]. However, other researchers have found no significant differences in Likert scales based on the presence of a mid-point alternative [22]. Another study on middle answer modalities [8] reveals that the reliability and validity of answers in a questionnaire measuring attitudes and opinions are not affected by the presence or absence of the middle option. Although the mid-point option seems to decrease the items' stability when considered individually, such a category allows for better psychometric coherence when the items are analysed jointly through latent dimensions. In Kulas et al. [20], two types of scales are used—a 5-point Likert scale with or without N/A (non-applicable response) option, and no significant difference in reliability and validity estimates was found. Nonetheless, the authors note that the overall frequency of N/A answers was low. Results of an eye-tracking study by Chen et al. [5] indicate that scales with five to seven response alternatives require the least cognitive effort. However, the five-point scales are the most desirable from an information-processing perspective, suggesting that the importance of the mid-point option on a scale seems to depend on the number of response alternatives. In fact, the neutral category is needed when the scale consists of a small number of alternatives (such as five). At the same time, it does not seem to exert a significant influence when a larger number of response categories (nine, for example) is provided.

Overall, there seems to be no consensus on whether to insert a mid-point category in agreement rating scales and in Likert-type scales. It appears that the choice of including or omitting the mid-point on a response scale depends on several factors whose effects are not yet well understood or confirmed. The prominence of Likert-type scales and their application to essential individual and societal decisions necessitates further insights into the potential benefits and drawbacks of using the mid-point, neutral category.

This paper examines the consequences of using the mid-point, neutral category, in research on non-cognitive skills. Specifically, our goal is to explore whether, in the pilot study on non-cognitive skills for the OECD's Programme for Assessment of Adult Competencies (PIAAC), such a category is used expectedly and validly to indicate the medium level of a measured attribute. We also aim to ascertain if the mid-point option improves or impairs the psychometric properties of various scales. Finally, we investigate which factors—related both to scales and respondents themselves—may influence the interviewees' tendency to employ the neutral response category validly.

The study is organized as follows. In Sect. 2 data are illustrated with a specific focus on the methodology implemented in the PIAAC pilot study. Descriptive statistics and findings related to the comparison of the psychometric properties of the scales across two experimental conditions are presented in Sect. 3. Section 4 is devoted to the discussion of the results. Some final remarks end the paper.

2 Data and methods

This paper employs data from the pilot study on non-cognitive skills developed within the research projects for the OECD's PIAAC. The Program aims to compare adult competencies across different countries for the sake of assessing the human capital of participating countries accurately and comprehensively while ensuring international comparability. The overall purpose of the PIAAC non-cognitive pilot study was to develop and test various non-cognitive scales that might be included in the main study [15, 16]. More detail on the study and its datasets can be found at GESIS Data Archive [27].

To determine whether the neutral category is being used as intended—specifically, to convey an intermediate level of the attribute to be assessed by a specific item, ensuring the psychometric properties of the scales, two forms of experimental survey designs were implemented:

- Between-subject experimental design—i.e., a split-ballot experiment—where two randomly selected groups of respondents are administered scales with and without a mid-point neutral category.
- Within-subject experimental design entailing the same respondents who are given two parallel forms of scales (in random order), also one with and one without the mid-point response category.

The survey questionnaire took around 20 min to complete (median time). Conducted entirely in English, it was carried out online from May to June 2015.

2.1 Sample

The experimental design of the pilot study considers two conditions (A and B), as detailed below. A quota sample design is implemented, and the participants are allocated between the 2 conditions. The sample consists of 2970 respondents in the between-subject design phase. In particular, we surveyed 2492 US residents, 1,193 of whom were selected for option A and 1299 for option B, and 478 UK residents, with 253 being assigned to option A and 225 to option B. Furthermore, additional 1606 US respondents were allocated to the within-subject design phase of the survey. Respondents were drawn from several online panels through a commercial contractor and were paid a commission for their participation in the survey.

The key variables, broadly representative of the US and UK census data, are respondents' gender, age, and regional distribution. However, the required quotas for gender were not fully achieved, since there are somewhat more women than men in both US and UK samples. Females are indeed 57.1% for condition A and 56.6% for condition B; married people are 45.3% and 47.3% for option A and B respectively. About 27.4% of option A respondents hold at least a high school degree; this percentage is about 23.8% for option B. For both options, 50% of the respondents are employed. The average age is 27.8 for respondents in option A and 27.5 in option B (13.3 is the standard deviation for both sample values). It is worth remarking that the slight differences in the characteristics of the two sub-samples in the between-group design—for all listed variables—have been checked for statistical significance and none of them was found to reach the level of significance ($p = 0.05$).

2.2 Instruments

The survey consists of five personality inventories measuring 12 personality constructs: the five Big Five dimensions, four Impulsivity dimensions, Traditionalism, Self-Efficacy [4] and Honesty/Integrity (one of the six HEXACO dimensions [2, 9]). The selected scales are either existing instruments, modifications of existing instruments (impulsivity scale) or a combination of existing instruments, namely, for self-efficacy and integrity/honesty. The main characteristics of the personality scales used in the survey are presented in Table 1.

Table 1 Measured constructs and related scales

Scale	Measured construct	Number of items	Labels of the response scale categories	Number of response categories
Big Five Inventory (BFI-2)[a]	Big Five	60	Agree/disagree	4 or 5
Chernyshenko Conscientiousness Scale[b]	Traditionalism	8	Agree/disagree	4 or 5
Short UPPS scale[c]	Impulsivity/Self-Control	20	Agree/disagree	4 or 5
General Self-Efficacy Short Scale[c] and The General Perceived Self-Efficacy Scale[d]	Self-efficacy	8	Agree/disagree	4 or 5
Brief HEXACO and IPIP-HEXACO[e]	Integrity-honesty	12	Agree/disagree	4 or 5

[a] Soto and John [31]

[b] Chernyshenko [6]

[c] Whiteside and Lynam [33]; Whiteside et al. [34]

[d] Beierlein et al. [4] and Schwarzer and Jerusalem [30] (Reduced version, Romppel et al. [29])

[e] de Vries [9] (Brief HEXACO) and Ashton et al. [2] (IPIP-HEXACO)

The questionnaire also comprises several socio-demographic, economic, and subjective well-being measures, as well as a short cognitive ability test [25]. Thorough information about the pilot study[1] and its datasets can be found at the GESIS data archive.[2]

Quality control criteria are as follows:

- Testing time—excluding those who answered in less than 6 min and marking those who had finished in between 6 and 8 min (1 point).
- Age—excluding those younger than 16 and older than 65.
- Ability test results (2 criteria)—marking those that repeatedly answered "don't know" or that did not have a single correct answer (1 point for each of the two criteria).
- Quality control items (3 criteria)—marking those who failed one, two or all three quality control answers (1 point for each failed quality control item).
- Consistency in answering.[3]

As mentioned, to ensure the quality of the information obtained, the survey contains control questions with only one correct answer: three quality control items[4] are placed within the personality scales to check for the quality of collected responses. These responses have been used, along with other data quality indicators, to create an overall quality control indicator (named "quality") and consequent exclusion of responses of poor quality.

The respondents excluded from our final analyses are those who:

- answered in less than 6 min;
- achieved a poor quality indicator (5 points or more);
- dropped out of the questionnaire or presented more than 11 non-responses.

2.3 Experimental design

The survey's main design factor is a variation of the presence of neutral/middle response options in response scales of the five personality scales. Two versions of each item response scale were utilized in the study: the first form included five response options, which encompassed the "neither agree nor disagree" option; the second one consisted of four response possibilities without the mid-point "neither agree nor disagree" option. These two versions yielded the following test conditions:

- Condition A: 5-point response scale with neutral mid-category;
- Condition B: 4-point response scale without neutral mid-category.

These two conditions were then examined using two different experimental designs:

- Between-subject design (phase 1);
- Within-subject design (phase 2).

In the first phase (between-subject design) each respondent was randomly assigned to one of the two test conditions. In addition to the five scales, which were presented in either a 4-point or 5-point response format, all other variables were presented to all respondents in the same format.

[1] https://doi.org/10.4232/1.13062.

[2] https://www.gesis.org/piaac/fdz/daten/piaac-pilot-studies-on-non-cognitive-skills.

[3] Consistency check was performed by marking those that gave the same answers to four pairs of opposing/reverse questions (e.g. "*Is neat*" vs. "*Is messy*"; or "*Is talkative*" vs. "*Tends to be quiet*").

[4] In particular, the three control items are: (a) *I live in the United States*; (b) *I am older than 16 years of age*; (c) *I fly to the International Space Station*.

The two phases have been carried out at consecutive times, with the between-subject design study preceding the within-subject design study a few weeks. The respondents were drawn from the same population of online respondents available in one of the world's largest commercial networks for digital survey-based research. Some of the respondents may have participated in both studies; nonetheless, we would not expect that such an occurrence has been prominent, thanks to the size of the online panels. However, we could not establish how many, if any, respondents participated in both phases of the experimental study.

During the second phase (within-subject design), three groups of respondents, consisting of 506, 552, and 548 respondents, respectively, were administered various personality scales under both test conditions. In particular, the first group of respondents was assigned the Big Five questionnaire under two different conditions. The second group was given Traditionalism, Impulsivity, and Self-Efficacy scales in both conditions. Integrity/Honesty was administered to the third group in both conditions. The OECD expert group decided in a few cases to modify the existing scales in a way that they deemed to be more fitting for their integration into the PIAAC [15, 27].

The two types of experimental designs are used to provide complementary evidence on the differences and similarities in psychometric properties of the selected scales across the two conditions. For example, the between-subject design allows for a comparison of test conditions without possible priming effects of test familiarity and with less cognitive load of having to answer the same questions two times. On the other hand, the within-subject design allows us to examine the responses in condition B of those individuals who selected the mid-point category in condition A.

3 Results

In this section, we first present descriptive statistics for all the scales across the two experimental conditions: condition A and condition B separately, and, for both experimental designs, between- and within-subject designs. The psychometric properties of the scales are then examined across both experimental conditions and designs. Factor analysis of individual scales is employed to compare the proportion of accounted variance by extracted first factors. Construct validity for the scales is then assessed through correlations with antecedent and outcome variables, as detailed in the following.

Within-subject data are used to present distributions of answers in condition B of those participants who chose the neutral category in condition A. Finally, an Item Response Theory (IRT) analysis on between-subject data is conducted to assess in which manner respondents have used the neutral response category [28].

3.1 Descriptive statistics

Means, medians, skewness, and kurtosis are computed for each scale in conditions A and B, for both between-subject and within-subject designs. To make the descriptives of the two scales comparable across the two conditions, we have coded the conditions as follows. In condition A, a 5-point agreement scale ranging from 1 to 5 is employed, with the neutral mid-point being 3 and thus representing the average of this scale. In condition B we adopt the following re-coding: 1—Strongly disagree, 2—Disagree, 4—Agree, 5—Strongly Agree. In this way, we achieve the same theoretical average (value equal to 3) across the two conditions. Indeed, if one would assume that those choosing a neutral category would equally split into categories 2 and 4 once faced with the scale used in condition B (the scale without a neutral category), then it should be expected that the average scores among the two conditions are entirely the same. Likewise, any differences in their average scores should be directly comparable and meaningful and could be related to the experimental conditions, given the random assignment of respondents to these conditions, and considering the lack of any differences in socio-demographic variables in the two samples.

The results, presented in Tables 2 and 3, display somewhat higher mean and median values of scales across the two conditions. Negative skewness is also slightly more pronounced in condition A. Moreover, although the average values of kurtosis are about the same, its absolute values are fairly higher in condition A. An analysis of statistical significance of mean scores across the two experimental conditions is performed for all scale items and the results of these analyses are added to the appendix in Table 9 (for the 60 Big-Five items) and 10 (for the 48 remaining items). Results indicate that, once reversed items are recoded, mean scores in the experimental condition A are smaller, for all the questions in the Big Five scales as well as for the majority of those in the other scales.

The effect sizes of item differences vary between zero to a maximum of 0.10, indicating a slight but consistent shift in one direction across the large majority of questions.

3.2 Comparison of psychometric properties of the scales across the two experimental conditions

To evaluate the reliability of each scale, we calculated Cronbach's alpha coefficients and summarized the results in Table 4 for all scales under different conditions and designs. The reliability measures are relatively high for most scales across all conditions, although the coefficients are only marginally higher in condition A as compared to condition B. The "cocron" package [10] for comparing the statistical significance of the difference in alpha coefficients of reliability allows us to display that several of the observed reliabilities are statistically significant. Interestingly, this slight increase in reliability across the two experimental conditions is more pronounced in the within-subject design.

Results of the factor analyses carried out for the 12 scales are presented in Tables 5 and 6. Oblimin factor rotation is utilized to account for mutual correlations of the four subscales of the Impulsivity scale: Urgency, Premeditation, Perseverance, and Sensation seeking. All other scales are treated as independent constructs and therefore are separately analysed. The percentage of variance explained by the first factor is then compared across the two conditions. Our findings display a substantially higher proportion of explained variance by the extracted first factors. This proportion is higher by approximately 3% in absolute values, implying an increase of about 8% in relative terms, in both between- and within-subject designs.

We have also examined scale correlations with variables representing related constructs, either as their possible antecedents or outcomes. In detail, as regards the antecedent variables we refer to gender, age, parents' education level, and immigration status. With respect to the outcome variables, we consider marital status, education, income, job satisfaction, self-assessed health, life satisfaction, and employment.

Correlations of the 12 scales with antecedent and outcome variables for conditions A and B using between-and within-subject data are presented in Tables 7 and 8, respectively. The statistical significance of these correlations was examined using the "cocron" procedure [11] and the significant differences between values for conditions A and B are highlighted.

Table 2 Descriptive statistics: between-subject design

	Condition A				Condition B			
	Mean	Median	Skewness	Kurtosis	Mean	Median	Skewness	Kurtosis
Big Five								
Extraversion	3.14	3.08	– 0.14	– 0.70	3.12	3.33	– 0.13	– 0.49
Agreeableness	3.75	3.83	– 0.68	0.52	3.53	3.50	– 0.45	0.03
Conscientiousness	3.80	4.00	– 0.69	0.19	3.56	3.50	– 0.50	– 0.03
Emotional regulation	2.79	2.58	0.24	– 0.67	2.83	2.75	0.23	– 0.39
Openness	3.63	3.92	– 0.56	– 0.16	3.48	3.42	– 0.44	– 0.23
Self-control								
Premeditation	3.95	4.00	– 0.72	0.83	3.63	3.50	– 0.42	0.46
Sensation seeking	2.80	2.80	0.07	– 0.86	2.89	2.90	0.07	– 0.68
Impulsivity	2.77	2.60	0.19	– 0.86	2.82	2.70	0.16	– 0.65
Persistence	3.64	3.80	– 0.58	0.23	3.48	3.50	– 0.39	0.18
Traditionalism	3.47	3.63	– 0.40	– 0.22	3.79	3.38	– 0.28	– 0.20
Self-efficacy	3.79	4.00	– 0.76	0.86	3.56	3.50	– 0.46	0.77
Integrity/Honesty	3.99	4.17	– 1.02	0.68	3.69	3.67	– 0.82	0.36
Average	3.46	3.53	– 0.42	– 0.01	3.37	3.30	– 0.29	– 0.07

Table 3 Descriptive statistics: Within-subject design

	Condition A				Condition B			
	Mean	Median	Skewness	Kurtosis	Mean	Median	Skewness	Kurtosis
Big Five								
Extraversion	3.20	3.33	– 0.18	– 0.73	3.12	3.25	– 0.13	– 0.38
Agreeableness	3.65	3.75	– 0.60	0.26	3.47	3.42	– 0.44	0.11
Conscientiousness	3.72	3.92	– 0.58	0.07	3.52	3.50	– 0.38	– 0.15
Emotional regulation	2.74	2.58	0.24	– 0.64	2.81	2.67	0.15	– 0.34
Openness	3.65	3.92	– 0.59	– 0.13	3.47	3.50	– 0.40	– 0.17
Self-Control								
Premeditation	3.95	4.00	– 0.79	1.08	3.61	3.50	– 0.31	0.89
Sensation seeking	2.85	2.80	0.00	– 0.93	2.87	2.70	0.02	– 0.62
Impulsivity	2.84	3.00	0.05	– 1.01	2.86	2.70	0.05	– 0.59
Persistence	3.65	3.80	– 0.68	0.28	3.43	3.50	– 0.46	0.61
Traditionalism	3.46	3.50	– 0.42	– 0.26	3.31	3.50	– 0.27	– 0.14
Self-efficacy	3.86	4.00	– 0.88	1.53	3.56	3.50	– 0.40	1.36
Integrity/Honesty	4.03	4.21	– 1.18	0.78	3.74	3.92	– 0.95	0.45
Average	3.47	3.57	– 0.47	0.03	3.31	3.31	– 0.29	0.09

Table 4 Reliability analysis: Cronbach's alpha coefficients for each scale across the two conditions

	Between-subject design		Within-subject design	
	Condition A	Condition B	Condition A	Condition B
Big Five				
Extraversion	0.85	0.85	**0.82**	**0.80**
Agreeableness	0.80	0.79	0.81	0.82
Conscientiousness	0.88	0.87	0.88	0.87
Emotional regulation	0.91	0.90	0.88	0.88
Openness	**0.83**	**0.81**	0.85	0.84
Impulsivity				
Premeditation	**0.79**	**0.76**	**0.84**	**0.80**
Sensation seeking	0.82	0.81	**0.86**	**0.84**
Urgency	0.87	0.87	0.91	0.90
Persistence	**0.74**	**0.70**	0.69	0.65
Traditionalism	0.78	0.76	0.74	0.71
Self-Efficacy	**0.92**	**0.91**	**0.92**	**0.89**
Integrity/Honesty	0.82	0.81	0.83	0.83
Average	0.83	0.82	0.84	0.82

Cronbach's alpha coefficients that are statistically significantly different at least at 5% across the two conditions are indicated in bold and underlined font

These results are largely in line with preliminary assumptions, indicating that scales with neutral response category tend to show higher reliability and slightly higher predictive validity. At the same time, the statistical significance of differences in percentages of the explained variance cannot be computed since the corresponding models are not nested. However, percentages of the explained variance themselves account for the effect size and, as such, they can be interpreted as a measure of the effect of experimental conditions.

Comparisons of correlations with antecedent variables disclose few substantial differences in the pattern of relationships between scale scores and other individual characteristics. Correlations with certain outcome variables also display a high degree of similarity, with slightly higher values for scales in condition A. This is particularly noticeable in their correlations with both job and life satisfaction variables.

3.3 Forced-choice behaviour of respondents choosing the neutral category

Our further aim was to examine the answers of those respondents who chose a neutral category in condition A, to the same question in forced-choice condition B, when there was no neutral category using our within-subject data. Distributions of the answers in condition B of respondents who chose the neutral option for the same questions in condition A are summarized in Fig. 1. The means presented in the plots are taken from condition A for each scale.

These findings indicate that more than 90% of the respondents who opted for the mid-point, neutral option in condition A have chosen one of the two mid-point options in condition B. The observed distribution of answers also indicates that when a stronger skewness towards

Table 5 Between-subject design: Proportion of accounted variance by extracted factors

	Percentage of variance explained by the first factor		Difference in % of the accounted variance in conditions A and B	The relative increase in % of the accounted variance in condition A compared to condition B
	Condition A	Condition B		
Big Five				
Extraversion	32.9	32.2	0.7	2.2%
Agreeableness	28.4	26.2	2.2	8.4%
Conscientiousness	39.0	36.3	2.7	7.4%
Emotional regulation	47.5	44.2	3.3	7.5%
Openness	30.6	27.5	3.1	11.3%
Impulsivity				
Premeditation	40.1	36.9	3.2	8.7%
Sensation seeking	47.6	46.5	1.1	2.4%
Urgency	53.9	54.8	− 0.9	− 1.6%
Persistence	27.7	18.3	9.4	51.4%
Traditionalism	34.3	31.2	3.1	9.9%
Self-Efficacy	60.6	55.1	5.5	10.0%
Integrity/Honesty	28.7	27.3	1.4	5.1%
Average for all scales	39.3	36.4	2.9	8.0%

the right end of the scale occurs (e.g., Self-efficacy), such a skewness is also reflected in the distribution of respondents' neutral answers. In other words, the more the scale is skewed, the more the distribution of neutral answers is skewed as well.

3.4 IRT results: category characteristic curves—scale level

For each of the 12 scales, we conducted several Item Response Theory (IRT) analyses, implementing Partial Credit Model specification for polytomous scale items. Our aim is to examine the psychometric properties of the individual answer categories for each of the 12 scales, focusing on the properties of the neutral answer category. Category characteristic curves (CCC) of our IRT models using the between-subject data in condition A are presented in Fig. 2.

The CCC are useful statistical indicators of the discriminant power and correct use of individual response answers. They represent the probability of a respondent endorsing a particular response category for a given level of respondent's latent trait. So, in the ordered, Likert-type scales, the CCC of different response options should be located in an equidistant, ordered structure with responses indicating low levels of latent trait (e.g. "strongly disagree") being located in the corresponding ends of the x-axes of the latent trait distribution and vice versa.

Table 6 Within-subject design: Proportion of accounted variance by extracted factors

	Percentage of variance explained by the first factor		Difference in % of the accounted variance in conditions A and B	The relative increase in % of the accounted variance in condition A compared to condition B
	Condition A	Condition B		
Big Five				
Extraversion	30.3	28.8	1.5	5.2%
Agreeableness	31.4	30.9	0.5	1.6%
Conscientiousness	39.5	36.5	3	8.2%
Emotional regulation	40.8	39.3	1.5	3.8%
Openness	35.1	33.4	1.7	5.1%
Impulsivity				
Premeditation	48.4	42.7	5.7	13.3%
Sensation seeking	56.8	52.3	4.5	8.6%
Urgency	64.8	63.4	1.4	2.2%
Persistence	21.1	14.4	6.7	46.5%
Traditionalism	33	30.3	2.7	8.9%
Self-Efficacy	58.1	51.6	6.5	12.6%
Integrity/Honesty	31.7	31.8	− 0.1	− 0.3%
Average for all scales	40.9	38.0	3.0	7.8%

The CCC of the neutral category should be distributed between those from the two adjoint response categories (“agree” and “disagree”) and, in the case of normally distributed answers, should present its peak around the middle of the latent trait distribution. In Fig. 2, the probability of selecting the neutral response category across different levels of the latent trait is shown by the purple curves.

Apart from CCC’s location, their shape—namely, the height of their peak and the broadness of their distribution—is an indicator of their discrimination power and, consequently, of their contribution to the overall reliability of the scale. In our 5-point agreeableness Likert scales, theoretical expectations would suggest that the CCC for the neutral category have a normal distribution with a mean close to 0. This distribution’s peak should also be equally distant and symmetrical from the peaks of the two contiguous response categories.

Finally, this distribution’s peak should be of the same height as the two adjoint response categories’ peaks. Obtained results vary substantially across the scales, though they generally indicate that the mid-point response category was largely used appropriately. As expected, results reveal a slight shift in the distribution of the neutral category curve to the left in those scales with a strong negative skewness. In all scales, neutral category curves are located in between the two adjoint category curves, in line with theoretical expectations. In most cases, they are roughly equally distanced from the adjoint distributions, and this symmetry is not preserved only for a few scales (e.g., Conscientiousness, Self-Efficacy). Results differ the most with respect to the height of the peak of the neutral CCC. The peaks of the CCC are

Table 7 Correlations of personality scales with antecedent variables across the two experimental conditions

Condition	Gender (Females)		Age		Education of father		Education of mother		Immigration status	
	A	B	A	B	A	B	A	B	A	B
Big Five										
Extraversion	– 0.06	– 0.04	– 0.01	– 0.01	0.06	0.12	0.06	0.12	0.18	0.17
Agreeableness	0.13	0.12	0.15	0.12	– 0.10	– 0.03	– 0.10	– 0.02	**– 0.18**	**– 0.06**
Conscientiousness	0.02	0.07	0.18	0.16	– 0.04	0.03	– 0.06	– 0.02	**0.01**	**0.16**
Emotional regulation	– 0.16	– 0.15	0.11	0.11	0.04	0.10	0.01	0.10	**0.01**	**0.18**
Openness	0.03	– 0.03	– 0.08	– 0.06	0.06	0.13	0.07	0.11	– 0.05	0.07
Impulsivity										
Premeditation	0.00	0.04	0.03	0.03	– 0.01	0.03	– 0.03	0.01	**0.02**	**0.27**
Sensation seeking	– 0.12	– 0.17	– 0.28	– 0.26	0.11	0.16	0.11	0.18	**0.19**	**0.11**
Urgency	0.06	0.02	– 0.11	– 0.10	– 0.02	– 0.08	– 0.01	– 0.05	**0.00**	**– 0.11**
Persistence	– 0.06	– 0.02	0.07	0.06	– 0.01	0.04	– 0.03	0.03	**0.01**	**0.15**
Traditionalism	0.05	0.09	0.11	0.06	– 0.10	– 0.05	– 0.10	– 0.05	0.04	0.08
Self-efficacy	– 0.06	– 0.07	0.01	0.03	0.06	0.07	0.03	0.06	**– 0.02**	**0.15**
Integrity/Honesty	0.16	0.20	0.18	0.19	– 0.15	– 0.10	– 0.13	– 0.11	**– 0.21**	**– 0.01**

Coefficients that are statistically significantly different at least at 5% across the two conditions A and B are indicated in underlined font, while those significant at 1% are in bold font

Table 8 Correlations of personality scales with outcome variables across the two experimental conditions

Condition	Married		Education		Income		Job satisfaction		Subjective health		Life satisfaction		Employment	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Big Five														
Extraversion	0.10	0.12	0.06	0.11	0.18	0.23	0.23	0.25	0.30	0.29	0.43	0.36	0.15	0.12
Agreeableness	0.08	0.02	– 0.04	– 0.04	0.04	0.04	0.15	0.17	0.07	0.03	0.26	0.22	0.02	– 0.03
Conscientiousness	0.16	0.11	0.06	0.09	0.20	0.17	**0.22**	**0.12**	0.21	0.23	0.32	0.29	0.13	0.09
Emotional regulation	14	0.07	0.13	0.16	0.21	0.23	0.27	0.31	0.34	0.35	0.53	0.52	0.20	0.14
Openness	– 0.05	– 0.04	0.12	0.12	0.06	0.07	0.09	0.11	0.13	0.11	0.18	0.14	0.01	– 0.02
Impulsivity														
Premeditation	0.02	0.04	0.04	0.04	0.08	0.06	**0.21**	**0.10**	0.08	0.15	0.17	0.21	0.02	0.03
Sensation seeking	– 0.03	– 0.01	0.10	0.12	0.08	0.12	**0.09**	**0.17**	0.26	0.26	0.16	0.17	0.11	0.08
Urgency	– 0.06	– 0.06	– 0.12	– 0.13	– 0.12	– 0.16	– 0.08	– 0.08	– 0.13	– 0.16	– 0.21	– 0.21	– 0.7	– 0.04
Persistence	0.08	0.05	0.07	0.06	0.15	0.14	0.24	0.20	0.21	0.25	0.36	0.31	0.10	0.07
Traditionalism	0.16	0.10	– 0.02	– 0.04	0.06	0.08	0.24	0.17	0.09	0.10	0.25	0.23	0.06	0.05
Self-efficacy	0.09	0.04	0.15	0.12	0.23	0.16	0.22	0.24	0.28	0.26	0.38	0.35	**0.21**	**0.09**
Integrity/Honesty	0.12	0.05	– 0.10	– 0.07	0.01	– 0.02	0.09	0.01	– 0.04	– 0.06	0.13	0.07	– 0.09	– 0.08

Coefficients that are statistically significantly different at least at 5% across the two conditions A and B are indicated in underlined font, while those significant at 1% are in bold font

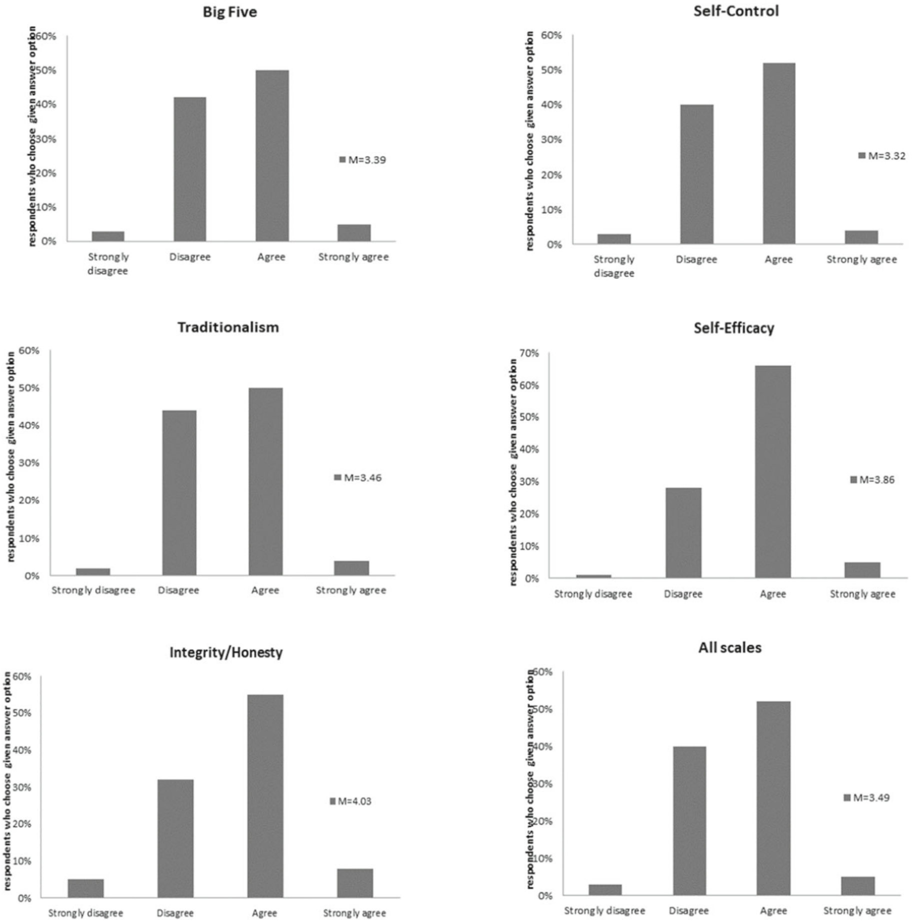

Fig. 1 Aggregated frequency distributions of answers in condition B from respondents who chose the neutral option on the same questions in condition A

similar to other response category CCC for some scales (e.g., Extraversion, Openness to Experience, Traditionalism, Persistence) and somewhat smaller than adjoint CCC for other scales (e.g., Agreeableness, Urgency, Integrity/Honesty).

4 Discussion

Taken together, the presented results indicate that, when the neutral category is used, a slight improvement in the psychometric properties of the selected personality scales occurs. In factor analyses, such an improvement is particularly noticeable in the proportion of variation explained by the first factor.

The observed increases in scales' reliability and proportion of accounted variation by the first factors are largely in line with findings by Lozano et al. [24] in their Monte Carlo simulation study of Likert scales with corresponding response options [24]. Such results

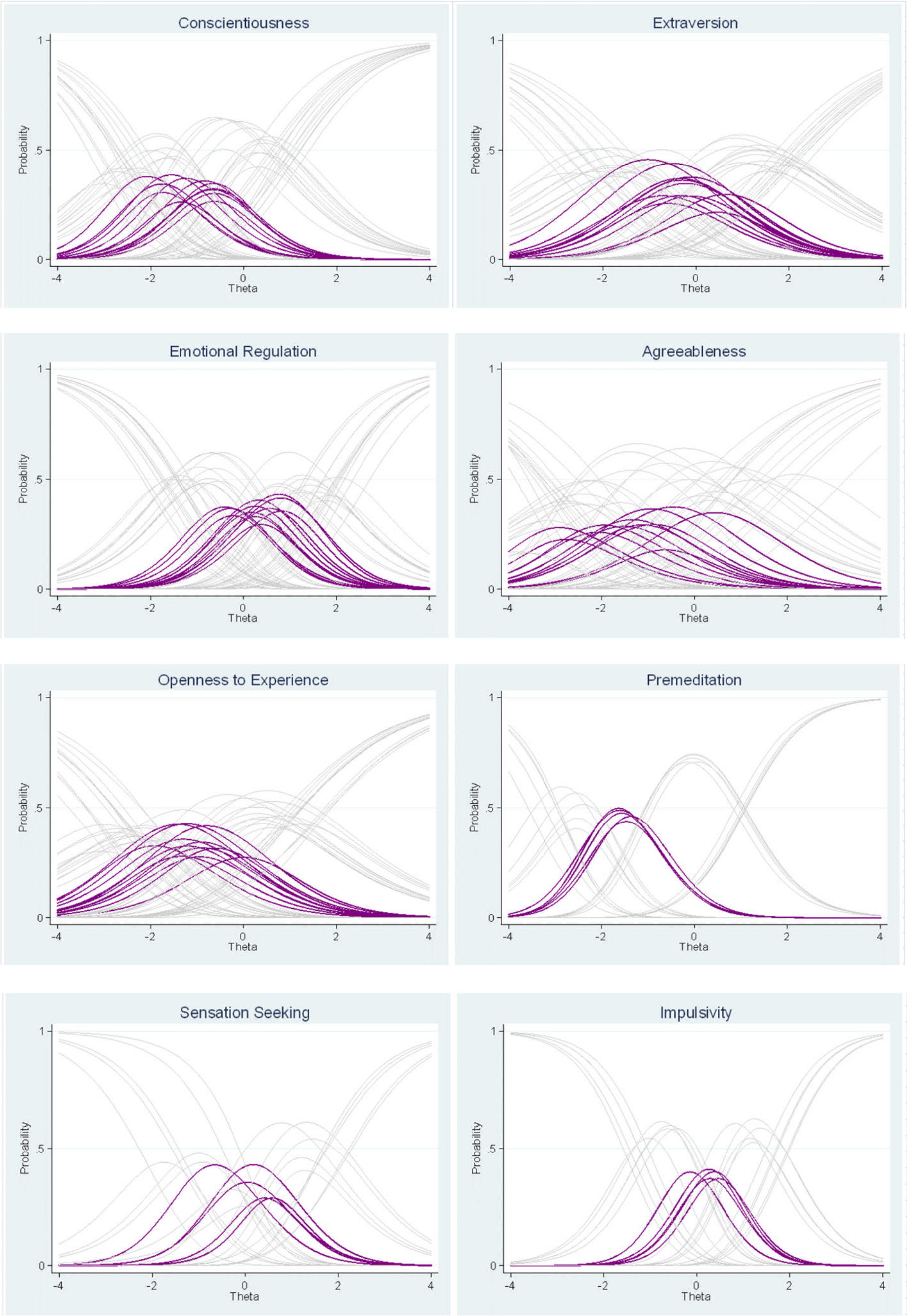

Fig. 2 IRT Category Characteristic Curves of the twelve personality scales with the neutral category response in the between-subject design

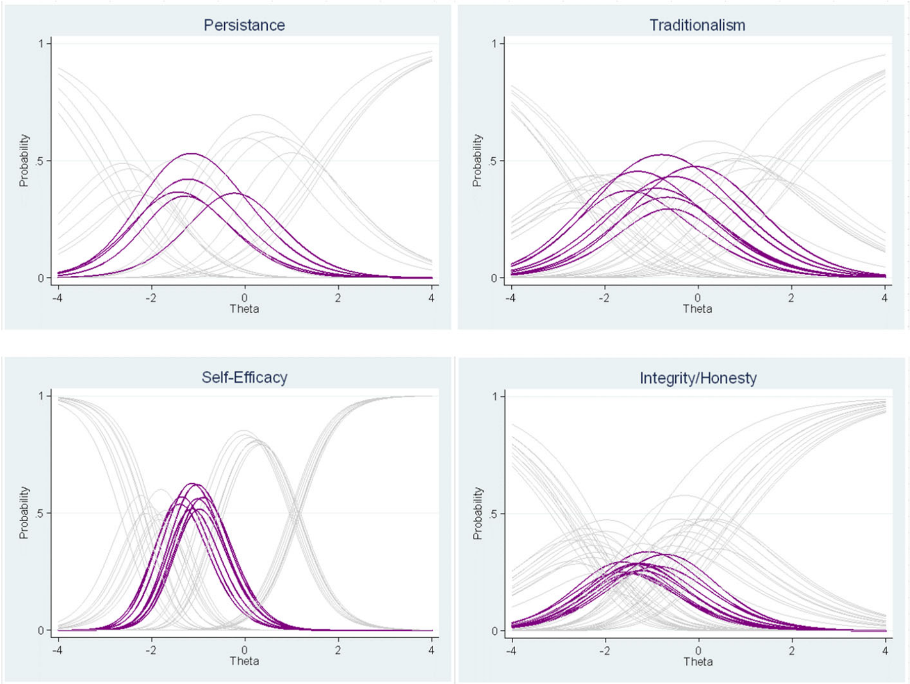

Fig. 2 continued

indicate higher coherence of information obtained from individual items with the neutral response category and lower measurement error level.

Although indicative, the descriptive analyses of the psychometric characteristics of the considered personality scales do not fully explain how respondents use the neutral category. The within-subject experimental data reveal that the large majority of the respondents who have chosen the neutral category, when it was available, were picking the contiguous response categories to "agree" or "disagree", when such an option was not offered. Such response behaviour is generally in line with theoretical expectations assuming valid responding attitudes. However, one should still note that a small proportion of interviewees have opted for more distant options indicating an invalid response pattern when choosing a neutral category in condition A or one of the two extreme response modalities in condition B.

The IRT findings offer a further indication that the mid-point, neutral category was used validly by the majority of respondents. The location of the neutral responses' category characteristic curves at the latent construct's distribution is always in between the two sets of contiguous response curves. In most cases, their location is also roughly equidistant from the two adjoint responses. Furthermore, in almost all of the scales, their category characteristic curves show information values similar to those of the adjacent response categories.

However, we could also notice that the use of mid-point scales somewhat varies across the scales and contains the highest amount of measurement noise in the case of Integrity/Honesty and Agreeableness scales. These results might show that in the case of these particular scales, at least a subset of respondents who were choosing the neutral response category were doing so for other reasons rather than to signify their true position on the given question. In interpreting these results, one should consider the relatively high social sensitivity of these two scales' concepts and behaviours. The Integrity/Honesty scale asks about respondents' tendency to lie, cheat, steal, etc., while the Agreeableness scale asks whether they treat others with respect or are argumentative, etc. In both cases, the assessed behaviours and dispositions imply socially normed and ethically charged issues, more deeply than the constructs and behaviours measured by other scales. Therefore, the fact that the neutral categories seem to be more invalidly used in these scales could be interpreted as an indication of respondents using the neutral modality as an "escape" category. In other words, at least some of the respondents could be using neutral options as a way to avoid providing answers that might be deemed as socially undesirable.

Our results indicate that the majority of respondents have been validly using the mid-point neutral category, i.e., as an indicator of their median position on a given question. This conclusion is supported by data from both between-subject and within-subject experiments. Such valid use of mid-point categories is shown to lead to slightly increased scale reliabilities and internal coherence in scales including the neutral response category. Within-subject data have further confirmed these results by showing largely valid answers from those respondents who chose the neutral category, when it was available, and asked the same question without neutral category present.

However, results also display that a minority of respondents might be using the neutral category in an invalid way, especially for some scales: up to 10% of respondents were choosing two extreme response options on a 4-point scale even though they had picked the neutral option on a 5-point scale. Even more illustratively, the information value of the neutral responses, as indicated by their category characteristic curves, in several scales is slightly lower (although still substantial) than the corresponding value of other response options. This is especially evident in the case of the Integrity/Honesty and Agreeableness scales. Such results could be interpreted in line with one of the identified possible invalid uses of the neutral response categories, i.e., a situation in which respondents are using a neutral response category to avoid providing socially undesirable answers to highly sensitive/socially normed questions.

5 Final remarks

Obtained results are in line with studies showing that both the reliability and validity of the scales are increased when more response options are given, suggesting that a mid-point option could be useful for improving the scale properties [7, 25]. Importantly, our findings also display that the mid-point category can also be employed in an invalid way. In our study, a minority of respondents seem to use it as an "escape" category when asked about socially sensitive behaviours. Such results might imply that this biasing threat would be more serious with scales measuring highly socially sensitive constructs and behaviours, which is in line with some previous findings on this topic [12, 31]. On the other hand, observed results do not indicate the presence of the other two potential invalid uses of the neutral category: satisficing (reducing cognitive costs) and hidden don't know [17, 18, 26].

Although providing a broad scope of empirical evidence and offering interesting insights into the neutral response category's workings, it is important to delineate some key limitations to this study's findings. First, the study was conducted using representative samples of UK and USA residents. Given that the use of the mid-point category may differ across different cultures our results do not necessarily hold beyond the two populations in our samples. Likewise, one should consider that results might vary also with other constructs measured especially in different response situations. For example, one might imagine that in high-stakes situations in which assessment results would lead to important personal consequences, respondents' use of the neutral response category could be rather different than that observed in our low-stakes research settings.

Appendix

See Tables 9 and 10.

Table 9 Sample tests' results for Big-Five items

Item	Scale	Mean difference	Std. error difference	*t*-test	*df*	Sig. (2-tailed)	Cohen's d
Is curious about many different things	Open-Mindedness	– 0.17	0.03	– 5.60	2968	0.00	0.10
Is reliable, can always be counted on	Conscientiousness	– 0.11	0.03	– 3.79	2968	0.00	0.07
Is original, comes up with new ideas	Open-Mindedness	– 0.14	0.04	– 3.61	2968	0.00	0.07
Is helpful and unselfish with others	Agreeableness	– 0.10	0.03	– 3.41	2968	0.00	0.06
Is persistent, works until the task is finished	Conscientiousness	– 0.11	0.03	– 3.40	2968	0.00	0.06
Has difficulty imagining things	Open-Mindedness	0.13	0.04	3.35	2968	0.00	0.06
Is efficient, gets things done	Conscientiousness	– 0.11	0.03	– 3.27	2968	0.00	0.06
Is respectful, treats others with respect	Agreeableness	– 0.08	0.02	– 3.25	2968	0.00	0.06
Is complex, a deep thinker	Open-Mindedness	– 0.12	0.04	– 3.17	2968	0.00	0.06
Is compassionate, has a soft heart	Agreeableness	– 0.09	0.03	– 3.10	2968	0.00	0.06
Is inventive, finds clever ways to do things	Open-Mindedness	– 0.11	0.04	– 2.92	2968	0.00	0.05
Values art and beauty	Open-Mindedness	– 0.11	0.04	– 2.88	2968	0.00	0.05
Keeps their emotions under control	Negative Emotionalitiy	– 0.10	0.04	– 2.61	2968	0.01	0.05
Prefers to have others take charge	Extraversion	0.11	0.04	2.60	2968	0.01	0.05
Has difficulty getting started on tasks	Conscientiousness	0.11	0.04	2.57	2968	0.01	0.05
Tends to find fault with others	Agreeableness	0.11	0.04	2.53	2968	0.01	0.05
Often feels sad	Negative Emotionalitiy	0.11	0.05	2.31	2968	0.02	0.04
Feels secure, comfortable with self	Negative Emotionalitiy	– 0.09	0.04	– 2.30	2968	0.02	0.04

Table 9 (continued)

Item	Scale	Mean difference	Std. error difference	*t*-test	*df*	Sig. (2-tailed)	Cohen's d
Has little interest in abstract ideas	Open-Mindedness	0.09	0.04	2.26	2968	0.02	0.04
Rarely feels excited or eager	Extraversion	0.09	0.04	2.19	2968	0.03	0.04
Avoids intellectual, philosophical discussions	Open-Mindedness	0.10	0.04	2.17	2968	0.03	0.04
Stays optimistic after experiencing a setback	Negative Emotionalitiy	– 0.09	0.04	– 2.14	2968	0.03	0.04
Finds it hard to influence people	Extraversion	0.09	0.04	2.04	2968	0.04	0.04
Is temperamental, gets emotional easily	Negative Emotionalitiy	0.09	0.05	1.98	2968	0.05	0.04
Has an assertive personality	Extraversion	– 0.09	0.05	– 1.96	2968	0.05	0.04
Is dependable, steady	Conscientiousness	– 0.05	0.03	– 1.72	2968	0.09	0.03
Is dominant, acts as a leader	Extraversion	– 0.08	0.05	– 1.65	2968	0.10	0.03
Tends to feel depressed, blue	Negative Emotionalitiy	0.08	0.05	1.65	2968	0.10	0.03
Has little creativity	Open-Mindedness	0.07	0.04	1.65	2968	0.10	0.03
Is fascinated by art, music, or literature	Open-Mindedness	– 0.07	0.04	– 1.63	2968	0.10	0.03
Is relaxed, handles stress well	Negative Emotionalitiy	– 0.07	0.04	– 1.63	2968	0.10	0.03
Worries a lot	Negative Emotionalitiy	0.07	0.05	1.49	2968	0.14	0.03
Thinks poetry and plays are boring	Open-Mindedness	0.07	0.05	1.48	2968	0.14	0.03
Is sometimes rude to others	Agreeableness	0.06	0.04	1.37	2968	0.17	0.03
Can be cold and uncaring	Agreeableness	0.06	0.04	1.29	2968	0.20	0.02
Shows a lot of enthusiasm	Extraversion	– 0.05	0.04	– 1.27	2968	0.20	0.02
Is emotionally stable, not easily upset	Negative Emotionalitiy	– 0.05	0.04	– 1.12	2968	0.26	0.02
Has a forgiving nature	Agreeableness	– 0.04	0.04	– 1.10	2968	0.27	0.02

Table 9 (continued)

Item	Scale	Mean difference	Std. error difference	*t*-test	*df*	Sig. (2-tailed)	Cohen’s d
Is full of energy	Extraversion	– 0.05	0.04	– 1.09	2968	0.28	0.02
Tends to be lazy	Conscientiousness	0.05	0.05	1.08	2968	0.28	0.02
Is moody, has up and down mood swings	Negative Emotionalitiy	0.05	0.05	1.07	2968	0.28	0.02
Is polite, courteous to others	Agreeableness	– 0.03	0.02	– 1.03	2968	0.31	0.02
Is systematic, likes to keep things in order	Conscientiousness	– 0.04	0.04	– 1.02	2968	0.31	0.02
Is less active than other people	Extraversion	0.05	0.05	1.00	2968	0.32	0.02
Rarely feels anxious or afraid	Negative Emotionalitiy	– 0.04	0.05	– 0.93	2968	0.35	0.02
Can be somewhat careless	Conscientiousness	0.04	0.04	0.85	2968	0.40	0.02
Tends to be disorganized	Conscientiousness	0.04	0.05	0.80	2968	0.42	0.01
Is sometimes shy, introverted	Extraversion	0.04	0.05	0.73	2968	0.47	0.01
Starts arguments with others	Agreeableness	0.03	0.04	0.71	2968	0.48	0.01
Feels little sympathy for others	Agreeableness	0.03	0.05	0.68	2968	0.50	0.01
Leaves a mess, doesn’t clean up	Conscientiousness	0.03	0.04	0.67	2968	0.50	0.01
Is outgoing, sociable	Extraversion	– 0.02	0.05	– 0.44	2968	0.66	0.01
Can be tense	Negative Emotionalitiy	0.02	0.04	0.39	2968	0.70	0.01
Tends to be quiet		0.01	0.05	0.31	2968	0.76	0.01
Sometimes behaves irresponsibly	Conscientiousness	– 0.01	0.04	– 0.20	2968	0.84	0.00
Keeps things neat and tidy	Conscientiousness	0.01	0.04	0.29	2968	0.77	0.01
Assumes the best about people	Agreeableness	0.01	0.04	0.32	2968	0.75	0.01
Is talkative	Extraversion	0.02	0.05	0.45	2968	0.65	0.01
Is suspicious of others’ intentions	Agreeableness	– 0.02	0.04	– 0.48	2968	0.63	0.01
Has few artistic interests	Open-Mindedness	– 0.03	0.05	– 0.65	2968	0.52	0.01

Table 10 Sample tests' results for items other than those of the BIG FIVE

Items	Scale	Mean difference	Std. error difference	*t*-test	*df*	*p* value	Cohen's d
When confronted with a problem, I do more than what is expected of me	Self-Efficacy	− 0.16	0.04	− 4.64	2968	0.00	0.09
No matter what comes my way, I'm usually able to handle it	Self-Efficacy	− 0.15	0.03	− 4.63	2968	0.00	0.08
Even difficult and complicated tasks I can successfully resolve	Self-Efficacy	− 0.14	0.03	− 4.49	2968	0.00	0.08
Thanks to my resourcefulness, I know how to handle unforeseen situations	Self-Efficacy	− 0.13	0.03	− 3.95	2968	0.00	0.07
Before making up my mind, I consider all the advantages and disadvantages	Self-Control	− 0.11	0.03	− 3.65	2968	0.00	0.07
I remain interested in the tasks that I start	Self-Control	− 0.12	0.03	− 3.61	2968	0.00	0.07
I can deal with most problems using my own resources	Self-Efficacy	− 0.10	0.03	− 3.55	2968	0.00	0.07
I continue working on tasks until everything is perfect	Self-Control	− 0.12	0.04	− 3.41	2968	0.00	0.06

Table 10 (continued)

Items	Scale	Mean difference	Std. error difference	*t*-test	*df*	*p* value	Cohen's d
When confronted with a problem, I give up easily	Self-Control	0.12	0.04	3.39	2968	0.00	0.06
In difficult situations I can rely on my skills	Self-Efficacy	– 0.10	0.03	– 3.33	2968	0.00	0.06
I am confident that I could deal efficiently with unexpected events	Self-Efficacy	– 0.11	0.04	– 3.11	2968	0.00	0.06
It is easy for me to stick to my aims and accomplish my goals	Self-Control	– 0.11	0.04	– 3.05	2968	0.00	0.06
People who resist authority should be severely punished	Traditionalism	0.13	0.04	2.98	2968	0.00	0.05
I can remain calm when facing difficulties because I can rely on my coping abilities	Self-Efficacy	– 0.11	0.04	– 2.96	2968	0.00	0.05
When I feel rejected, I will often say things that I later regret	Self-Control	0.12	0.05	2.54	2968	0.01	0.05
I would enjoy parachute jumping	Self-Control	– 0.13	0.05	– 2.50	2968	0.01	0.05
In my opinion, all laws should be strictly enforced	Traditionalism	0.10	0.04	2.31	2968	0.02	0.04

Table 10 (continued)

Items	Scale	Mean difference	Std. error difference	*t*-test	*df*	*p* value	Cohen's d
I usually make up my mind through careful reasoning	Self-Control	– 0.07	0.03	– 2.27	2968	0.02	0.04
I usually think carefully before doing anything	Self-Control	– 0.06	0.03	– 1.99	2968	0.05	0.04
I welcome new and exciting experiences and sensations, even if they are a little frightening and unconventional	Self-Control	– 0.08	0.04	– 1.92	2968	0.05	0.04
I sometimes like doing things that are a bit frightening	Self-Control	– 0.08	0.05	– 1.82	2968	0.07	0.03
When I am upset I often act without thinking	Self-Control	0.07	0.05	1.52	2968	0.13	0.03
I use flattery to get ahead	Integrity-Honesty	0.06	0.04	1.49	2968	0.14	0.03
I put off difficult problems	Self-Control	0.07	0.04	1.48	2968	0.14	0.03
People respect authority more than they should	Traditionalism	0.06	0.04	1.46	2968	0.14	0.03
I tend to value and follow a rational, "sensible" approach to things	Self-Control	– 0.04	0.03	– 1.45	2968	0.15	0.03
I pretend to be concerned for others	Integrity-Honesty	0.06	0.04	1.43	2968	0.15	0.03

Table 10 (continued)

Items	Scale	Mean difference	Std. error difference	*t*-test	*df*	*p* value	Cohen's d
I act like different people in different situations	Integrity-Honesty	0.07	0.05	1.41	2968	0.16	0.03
I would enjoy the sensation of skiing very fast down a high mountain slope	Self-Control	− 0.07	0.05	− 1.27	2968	0.20	0.02
I support long-established rules and traditions	Traditionalism	− 0.05	0.04	− 1.23	2968	0.22	0.02
I have the highest respect for authorities and assist them whenever I can	Traditionalism	− 0.03	0.04	− 0.97	2968	0.33	0.02
I would never take things that aren't mine	Integrity-Honesty	− 0.03	0.04	− 0.95	2968	0.35	0.02
I would like to know how to make lots of money in a dishonest manner	Integrity-Honesty	0.03	0.04	0.87	2968	0.39	0.02
I believe that people should be allowed to take drugs, as long as it doesn't affect others	Traditionalism	− 0.04	0.05	− 0.81	2968	0.42	0.01
In the heat of an argument, I will often say things that I later regret	Self-Control	0.04	0.05	0.80	2968	0.43	0.01
I am a cautious person	Self-Control	− 0.02	0.03	− 0.72	2968	0.47	0.01

Table 10 (continued)

Items	Scale	Mean difference	Std. error difference	*t*-test	*df*	*p* value	Cohen's d
When working with others I am the one who makes sure that rules are observed	Traditionalism	− 0.03	0.04	− 0.67	2968	0.51	0.01
I would never cheat on my taxes	Integrity-Honesty	− 0.02	0.04	− 0.58	2968	0.56	0.01
I don't pretend to be more than I am	Integrity-Honesty	− 0.02	0.04	− 0.48	2968	0.63	0.01
I find it difficult to lie	Integrity-Honesty	− 0.02	0.04	− 0.45	2968	0.65	0.01
Even if I knew how to get around the rules without breaking them, I would not do it	Traditionalism	0.02	0.04	0.42	2968	0.67	0.01
I return extra change when a cashier makes a mistake	Integrity-Honesty	0.02	0.04	0.37	2968	0.71	0.01
I often make matters worse because I act without thinking when I am upset	Self-Control	− 0.01	0.05	− 0.24	2968	0.81	0.00
I quite enjoy taking risks	Self-Control	0.01	0.05	0.18	2968	0.85	0.00
Sometimes I do impulsive things that I later regret	Self-Control	0.01	0.05	0.16	2968	0.88	0.00
I admire a really clever scam	Integrity-honesty	0.00	0.04	0.07	2968	0.94	0.00
I put on a show to impress people	Integrity-honesty	0.00	0.04	− 0.05	2968	0.96	0.00
I cheat on people who have trusted me	Integrity-honesty	0.00	0.03	0.00	2968	1.00	0.00

Author contributions All authors equally contributed to the study and are responsible for its content and they have agreed to this submission. All authors read and approved the final manuscript; The manuscript has not been previously published in any language anywhere.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript; The authors have no relevant financial or non-financial interests to disclose.

Data availability Data are available at GESIS data repository. For more detail, visit: https://www.gesis.org/en/piaac/research-data-center-piaac. Data may be requested from: https://search.gesis.org/research_data/ZA6940?doi=10.4232/1.13062.

References

1. Allen, I.E., Seaman, C.A.: Likert scales and data analyses. Qual. Prog. **40**, 64–65 (2007)
2. Ashton, M.C., Lee, K., Goldberg, L.R.: The IPIP-HEXACO scales: an alternative, public-domain measure of the personality constructs in the HEXACO model. Personal. Ind. Differ. **42**(8), 1515–1526 (2007). https://doi.org/10.1016/j.paid.2006.10.027
3. Baka, A., Figgou, L., Triga, V.: "Neither agree, nor disagree": a critical analysis of the middle answer category in Voting Advice Applications. Int. J. Electron. Gov. **5**(3–4), 244–263 (2013). https://doi.org/10.1504/IJEG.2012.051306
4. Beierlein, C., Kemper, C.J., Kovaleva, A., Rammstedt, B.: Short scale for measuring general self-efficacy beliefs (ASKU). Methods Data Anal. **7**(2), 251–278 (2013). https://doi.org/10.12758/mda.2013.014
5. Chen, X., Yu, H., Yu, F.: What is the optimal number of response alternatives for rating scales? From an information processing perspective. J. Mark. Anal. **3**(2), 69–78 (2015). https://doi.org/10.1057/jma.2015.4
6. Chernyshenko, O.S.: Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness. Ph.D. Dissertation, University of Illinois at Urbana-Champaign (2002)
7. Courtenay, B.C., Weidemann, C.: The effects of a "don't know" response on Palmore's facts on aging quizzes. Gerontologist **25**(2), 177–181 (1985). https://doi.org/10.1093/geront/25.2.177
8. Dassa, C., Lambert, J., Blais, R., Potvin, D., Gauthier, N.: Effects of a neutral answer choice on the reliability and validity of attitude and opinion items. Can. J. Prog. Eval. **12**(2), 61–80 (1997)
9. de Vries, R.E.: The 24-item Brief HEXACO Inventory (BHI). J. Res. Personal. **47**(6), 871–880 (2013). https://doi.org/10.1016/j.jrp.2013.09.003
10. Diedenhofen, B., Musch, J.: cocor: a comprehensive solution for the statistical comparison of correlations. PLoS ONE **10**, e0121945 (2015). https://doi.org/10.1371/journal.pone.0121945
11. Diedenhofen, B., Musch, J.: cocron: a Web Interface and R Package for the statistical comparison of Cronbach's alpha coefficients. Int. J. Internet Sci. **11**(1), 51–60 (2015)
12. Garland, R.: The mid-point on a rating scale: Is it desirable? Mark. Bull. **2**(1), 66–70 (1991)
13. Jamieson, S.: Likert scales: How to (ab)use them. Med. Educ. **38**(12), 1217–1218 (2004). https://doi.org/10.1111/j.1365-2929.2004.02012.x
14. Jebb, A.T., Ng, V., Tay, L.: A review of key Likert Scale development advances: 1995–2019. Front. Psychol. **12**, 637547 (2021). https://doi.org/10.3389/fpsyg.2021.637547
15. Kankaraš, M.: Personality Matters: Relevance and Assessment of Personality Characteristics, OECD Education Working Papers, No. 157, OECD Publishing, Paris. https://doi.org/10.1787/8a294376-en (2017)
16. Kankaraš, M. Suarez-Alvarez, J.: Assessment framework of the OECD Study on Social and Emotional Skills. In: OECD Education Working Papers, 207, OECD Publishing, Paris. https://doi.org/10.1787/5007adef-en (2019)

17. Krosnick, J.A.: Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl. Cogn. Psychol. **5**(3), 213–236 (1991). https://doi.org/10.1002/acp.2350050305
18. Krosnick, J.A., Narayan, S., Smith, W.R.: Satisficing in surveys: initial evidence. New Dir. Eval. **70**, 29–44 (1996). https://doi.org/10.1002/ev.1033
19. Kulas, J.T., Stachowski, A.A.: Middle category endorsement in odd-numbered Likert response scales: associated item characteristics, cognitive demands, and preferred meanings. J. Res. Personal. **43**(3), 489–493 (2009). https://doi.org/10.1016/j.jrp.2008.12.005
20. Kulas, J.T., Stachowski, A.A., Haynes, B.A.: Middle response functioning in Likert-responses to personality items. J. Bus. Psychol. **22**(3), 251–259 (2008). https://doi.org/10.1007/s10869-008-9064-2
21. Lam, T.C., Allen, G., Green, K.E.: Is "neutral" on a Likert scale the same as "Don't know" for informed and uninformed respondents? Effects of serial position and labeling on selection of response options. Ontario Institute for Studies in Education of the University of Toronto (OISE/UT), Toronto, Canada (2010)
22. Leung, S.O.: A comparison of psychometric properties and normality in 4-, 5-, 6- and 11-point Likert scales. J. Soc. Serv. Res. **37**(4), 412–421 (2011). https://doi.org/10.1080/01488376.2011.580697
23. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **22**(140), 1–55 (1932)
24. Lozano, L.M., García-Cueto, E., Muñiz, J.: Effect of the number of response categories on the reliability and validity of rating scales. Methodol. Eur. J. Res. Methods Behav. Soc. Sci. **4**(2), 73–79 (2008). https://doi.org/10.1027/1614-2241.4.2.73
25. Madden, T.M., Klopfer, F.J.: The "cannot decide" option in Thurstone-type attitude scales. Educ. Psychol. Meas. **38**(2), 259–264 (1978). https://doi.org/10.1177/001316447803800207
26. Nadler, J.T., Weston, R., Voyles, E.C.: Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. J. Gener. Psychol. **142**(2), 71–89 (2015). https://doi.org/10.1080/00221309.2014.994590
27. Organisation for Economic Co-operation and Development (OECD): Programme for the International Assessment of Adult Competencies (PIAAC), English Pilot Study on Non-Cognitive Skills. GESIS Data Archive, Cologne. ZA6940 Data file Version 1.0.0. https://doi.org/10.4232/1.13062(2018).
28. Reise, S.P., Widaman, K.F., Pugh, R.H.: Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol. Bull. **114**(3), 552 (1993)
29. Romppel, M., Herrmann-Lingen, C., Wachter, R., Edelmann, F., Düngen, H.D., Pieske, B., Grande, G.: A short form of the general self-efficacy scale (GSE-6): development, psychometric properties, and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc. Med. **10**, 1 (2013). https://doi.org/10.3205/psm000091
30. Schwarzer, R., Jerusalem, M.: Measures in health psychology: a user's portfolio. Causal and Control Beliefs **1**, 35–37 (1995). https://doi.org/10.1037/t00393-000
31. Soto, C., John, O.P.: The next Big Five Inventory (BFI-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J. Personal. Soc. Psychol. **113**(1), 117–143 (2017). https://doi.org/10.1037/pspp0000096
32. Sturgis, P., Roberts, C., Smith, P.: Middle alternatives revisited: how the neither/nor response acts as a way of saying "I don't know"? Sociol. Methods Res. **43**(1), 15–38 (2014). https://doi.org/10.1177/0049124112452527
33. Whiteside, S.P., Lynam, D.R.: The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. Personal. Ind. Differ. **30**(4), 669–689 (2001). https://doi.org/10.1016/S0191-8869(00)00064-7
34. Whiteside, S.P., Lynam, D.R., Miller, J., Reynolds, S.: Validation of the UPPS impulsive behavior scale: a four-factor model of impulsivity. Eur. J. Personal. **19**, 559–574 (2005). https://doi.org/10.1002/per.556